PSS718 - Data Mining
Assignment 1

due 26 Oct 2016

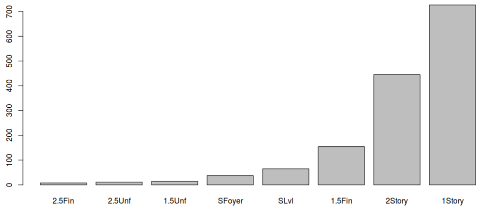
Follow the steps below. When you see R provide R code. When you see V provide visualizations. When you see E provide explanation in text. When you see a figure along the question, that is provided as a guide and an example. You may try to replicate the figure as much as possible but you should be answering the question using the figure you have generated on your own.

To answer some of the following, you may need to create your own supplementary data sets. When that is so, also provide the R code for creating those sets (or a CSV file output).

You may prepare your report in Latex or Word, I will accept only hard copies, and I will accept hard copies only if everything is clearly readable.

Try to do/answer the following:

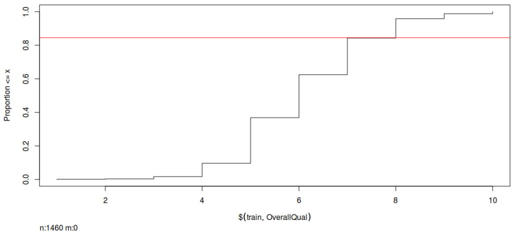
1. Download the *train.csv* and *desc.txt* files. *train.csv* contains the training data, and *desc.txt* contains explanations for variables and values
2. Read the training data from the CSV file
3. Show a comparison of the frequencies of the house styles V



4. What are the top three most frequently sold MSSubClasses? Your answer must be in text and not numbers. (So, ”SPLIT FOYER” is accepted but ”85” is not) R

```
[1] "1-STORY 1946 & NEWER ALL STYLES"
[1] "2-STORY 1946 & NEWER"
[1] "1-1/2 STORY FINISHED ALL AGES"
```

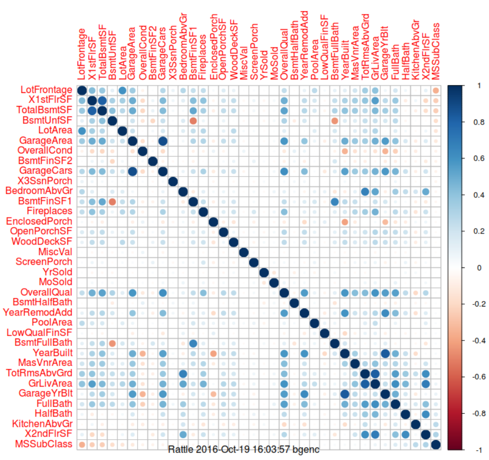
5. In terms of Overall Quality, what is the probability (roughly) that a house’s overall quality is 8 or more? E V



6. How many houses have been remodeled or had additions? What is the percentage of remodelling? How does the percentage change for houses built before and after (including) 2000? R

7. Do a correlation analysis. What do you see? E V

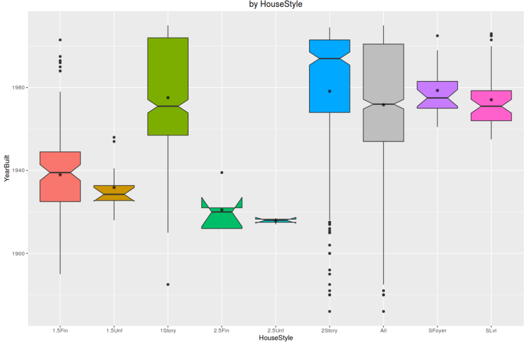
Correlation train.csv using Pearson



Rattle 2016-Oct-19 16:03:57 bgenc

8. How did the house styles change in time? E V

Distribution of YearBuilt (sample)
by HouseStyle



Rattle 2016-Oct-19 16:29:37 bgenc